Retrieval Feedback for Query Design in MEDLINE. A Comparison with Expert Network and LLSF Approaches

Padmini Srinivasan, Ph.D.
School of Library and Information Science
The University of Iowa, Iowa City, IA 52242

Query design is a vibrant research focus in information retrieval. The objective is to modify a user's original query into one that is more effective for retrieval. Researchers have proposed and investigated a variety of strategies to support query design in MEDLINE. This paper examines a query design method built within the framework of retrieval feedback. In particular, the effectiveness of this method is compared with the effectiveness of competing methods that utilize query mapping functions constructed in an expert network or built using the Linear Least Squares Fit approach. The comparison indicates that retrieval feedback offers an approach that is just as effective as these alternative approaches. Moreover, it has the advantage of being simpler to implement.

INTRODUCTION

The objective in query design is to automatically modify a user's original query into one that is more effective for retrieval. Strategies supporting this function have been built using both semantic [1, 2, 3] and statistical approaches [4, 5, 6, 7, 8].

In MEDLINE, query formulation strategies typically operate by adding query terms appropriately selected from MEDLINE's free–text, MeSH and/or the UMLS Metathesaurus vocabularies [9]. Hersh et al. utilize their experimental system SAPHIRE to identify UMLS Metathesaurus concepts [1]. Aronson et al., from the National Library of Medicine, explore a method based on syntactic analysis [2, 3] with the same vocabulary. Their method is similar to the one proposed and investigated by Elkin et al. [10].

Yang and Chute use mapping functions to select MeSH concepts given query words, where the mapping functions are learned from a training MEDLINE database [4, 5, 6]. Their work is a part of a larger investigation to solve for vocabulary differences between the queries and the documents. In [6] these mappings are derived using LLSF while in [4] an Expert Network is used. In either case their method requires a training set of example queries and their relevant documents. When applied to an 1,820 documents subset of the MEDLINE collection [1], Yang and Chute observe a 32.2% improvement in average precision over their baseline precision score of 0.412 [4, 6].

A practical disadvantage of the alternative Yang and Chute approaches are their reliance on the availability of relevance information for training, which limits the applicability of their method. More importantly, since 88% of their test queries appear in their training set they make the debatable assumption that a new query is likely to be similar to at least some of the queries in the training set. Since performance data for the non repeating 12% of test queries are not provided, it is not possible to predict what might happen in realistic situations when this assumption will most likely be violated.

In recent papers [11, 12] we investigated the potential of retrieval feedback as an overall framework for automated query design. Retrieval feedback is a derivative of relevance feedback [13, 14] with the difference that documents used for feedback information need not be actually relevant to the user [11, 15]. In retrieval feedback based query formulation a user's original query is used to conduct an initial retrieval run on the document collection. Information from the top few documents retrieved by this initial run is used to modify the original query. Since user relevance judgments are not required there is no additional cognitive responsibility placed on the user. Moreover there is no reliance on previous users of the retrieval system as is the case with a training set of relevance judgments. Consequently, no implicit assumption is being made about the relationship of the current query to the previous set of queries faced by the system. In terms of efficiency too, retrieval feedback has the advantage. It does not require the

0195—4210/96/$5.00 © 1996 AMIA, Inc.

creation of special mapping functions or networks. Such devices may need to be re–computed as the database grows and evolves over time. In contrast, retrieval feedback offers an approach that is relatively more stable.

In [11, 12] we tested this query expansion method using a MEDLINE database of 2,344 documents [1] and the Smart retrieval system [16, 17]. In brief, the best result obtained using queries expanded via retrieval feedback was a 11–AvgP (described later) score of 0.6018. This represents a statistically significant 16.4% improvement over a baseline 11–AvgP score of 0.5169 obtained with unexpanded queries. These papers also show how the improvements in retrieval performance are superior to those obtained by the alternative methods explored by [1, 2, 3].

Unfortunately, due to significant differences in experimental design we were unable to compare our previous results to those obtained by Yang and Chute [4, 5, 6] using their Expert Network and Linear Least Squares Fit (LLSF) methods. To explain, although Yang and Chute use the same 2,344 document MEDLINE database [1] for their experiments, they test their technique only on a subset of 1,820 documents. (They use the remaining 532 documents as a training set to create their mappings or expert network). In contrast, our experiments were run on the full set of 2,344 documents. More critically, less than 25% of the documents in their test database are relevant to at least one test query while this proportion is greater than 40% in the full set used by us. It may be that this difference in relevance proportion biases the results in favour of the retrieval feedback approach.

Thus our current goal is to conduct an experiment designed to support a comparison with the Yang and Chute results. The general objective is to study the relative merits of these alternative approaches to MEDLINE query design.

METHODS

This experiment testing retrieval feedback for query design was run using Cornell's SMART retrieval system [16, 17], a sophisticated and powerful research system based on the vector space model and designed for testing ideas pertaining to information retrieval. Comparative data for the LLSF and Expert Network approaches were obtained from [4, 6].

MEDLINE Test Collection

The test collection of 75 queries and 2,344 MEDLINE documents produced by Hersh et al. [1] is again used for this study. This collection includes all documents with abstracts, retrieved for the 75 queries. The queries and relevance judgments were specified by clinicians. Finally, all queries have some relevant documents in the collection.

Yang and Chute split the collection of 2,344 documents and 75 queries into a training set and a test set [5]. Our first goal is to reproduce the test set they created. The 2,344 documents contain 991 documents (42%) relevant to at least 1 of the 75 queries leaving 1,353 documents (58%) that are not relevant to any queries. Their procedure starts by sorting the 1,074 relevant query/document pairs by document number. Next, documents in the odd pairs form the training set while the remainder form the test dataset. They state that their resultant training set contains 524 relevant documents and their associated 71 queries. Therefore their test set has 1,820 documents and 68 queries. However, although we start with the same set of 1,074 relevant document/query pairs, when we reproduce their partitioning procedure we get 537 documents in the training set*. This gives us 1,807 (2,344 – 537) documents and their corresponding 65 queries for the test set. We use this test set for our retrieval feedback based query expansion experiment.

Indexing Strategies

In SMART, documents and queries are automatically indexed to yield a weighted vector of index terms. Term weights reflect the relative importance of terms when representing the textual unit. Details of the indexing strategy are identical to those described in [11, 12]. Only a brief summary is given here due to space constraints. Two word–based index vectors are derived for each document. A vector from the non trivial words in the title and abstract (ta–vector) and a vector from the non trivial words of the MeSH concepts (m–vector). We generate a single ta–vector from the free-text

*The 1,074 relevant document/query pairs, consist of 908 documents relevant to only one query, and 83 relevant to two queries. Thus given that they select every odd row from the list sorted by document number, their method guarantees that duplicate documents will not be selected. Thus the number in the training set should equal 537, i.e., exactly half of the number of relevant document/query pairs.

ta–vector Term	ta–vector Weight
option	0.69403
fungoid	0.46915
mycos	0.46158
assess	0.26772
treat	0.10234
patient	0.05497

Table 1: Original Query Representation: ta–vector.

of each query. The retrieval feedback process in SMART adds an m–vector to each query.

Since both documents and queries are represented by weighted vectors, SMART conducts retrieval by computing the similarity as the vector inner product of the document and the query vectors. Thus every query–document pair yields a numerical similarity value representing the closeness between the two entities. SMART uses these similarity values to rank the entire database for a given query. (This is in contrast to standard Boolean retrieval systems which partition the database into two sets for a given query: documents that are retrieved and documents that are not retrieved.) Given that SMART ranks all documents of the database by query similarity, the retrieved results, i.e., items shown to the user, consists of all documents above a threshold rank or a threshold similarity value. This threshold may be set by the user.

An Example of Query Expansion

Original Text of User Query: "Patient with mycosis fungoides, wishes to assess treatment options."

Table 1 shows the original ta–vector representation generated by SMART for the query. Note that the query words have been stemmed as part of the indexing process. This query is used to conduct an initial retrieval run. That is, documents of the database are ranked by their similarity to this initial query representation. The top few documents (in this example, top 10 documents) forms the feedback set. Finally a pre–defined number of MeSH terms (in this case, 10 terms) are selected from the feedback documents to create an m–vector for the query as shown in Table 2. Note that the overlap between entries in the two vectors is low; only two terms are in common: mycos and fungoid.

ta–vector Term	ta–vector Weight	m–vector Term	m–vector Weight
fungoid	0.46915	fungoid	0.32915
mycos	0.46158	mycos	0.31569
treat	0.10234	drug	0.07594
option	0.69403	human	0.01229
patient	0.05497	skin	0.26426
assess	0.26772	neoplasm	0.17483
		therap	0.07904
		age	0.06999
		middl	0.06551
		male	0.05348

Table 2: Final Query Representation with an additional m–vector.

Evaluation of Retrieval Strategies

Since SMART retrieves documents by ranking them against queries, alternative retrieval strategies are compared on their ranking effectiveness, i.e., their ability to rank relevant documents in the database higher than non relevant ones. The standard 11 point average precision (11–AvgP) measure is designed to evaluate ranked sets of documents. Recall is the proportion of relevant documents retrieved while precision is the proportion of retrieved documents that is relevant. Given a ranked set of documents, precision may be computed at the 11 standard recall points of 0%, 10%, ..., 100%. The final precision score of a retrieval strategy at a standard recall point is the average of precision scores at that point computed for each test query. This averaging technique yields *macro average* data wherein each test query is allowed to contribute equally to the overall performance score for the system (page 538 [17]). It should be noted that Yang and Chute ignore the precision score at the standard recall point of 0.0. Thus they compute 10 point average precision scores (10–AvgP).

RESULTS

Column 2 of Table 3 presents our baseline precision scores, at the 11 standard recall points of column 1, using the original user queries. Column 3 provides precision scores after the queries have been expanded using retrieval feedback. Thus a 19% and 21% improvement in 11–AvgP and

Recall	Baseline Precision	Retrieval Feedback Precision	Yang & Chute Precision
0.00	0.6485	0.6912	na
0.10	0.6190	0.6777	0.69
0.20	0.5761	0.6476	0.67
0.30	0.5506	0.6275	0.65
0.40	0.5199	0.5984	0.61
0.50	0.5108	0.5846	0.59
0.60	0.4260	0.5101	0.53
0.70	0.3765	0.4581	0.48
0.80	0.3219	0.4302	0.42
0.90	0.2825	0.4896	0.39
1.00	0.2626	0.3920	0.33
11–AvgP	0.4631	0.5517	na
10–AvgP	0.446	0.5416	0.536

Table 3: Precision at 11 Standard Recall Points. na = not available.

10–AvgP scores respectively may be observed. These represent statistically significant improvements ($p<0.01$) when tested using the non parametric Wilcoxon signed–rank test for matched samples. The improvements were achieved with a feedback set of 10 documents and by creating an m–vector of 10 terms for each query[†].

Column 4 of Table 3 presents the retrieval performance achieved by Yang and Chute. This data has been taken from [6] where they test the LLSF method. Yang achieves almost identical performance using their Expert Network (10–AvgP = 0.545) [4]. However, we were unable to obtain the individual precision scores for this method from the published literature. They compare these scores against a baseline precision score of 0.412. Interestingly their baseline, which is 7.6% lower than our baseline for the same database, was also achieved using the SMART system. The difference may be explained by differences in the SMART indexing parameters used.

These data indicate that their best performance, obtained using the Expert Network offers less than a 1% improvement over query design via retrieval feedback. Thus the two alternative methods for query design produce equivalent retrieval results. Retrieval feedback offers the same improvements without any investment in building expert networks or executing algorithms based on LLSF method. More importantly, retrieval feedback operates independent of any compilations of user relevance decisions. Interestingly, the 19% improvement in 11-AvgP score obtained in this experiments is almost identical to the 16.4% obtained previously using the full set of 2,344 documents MEDLINE database [12]. Thus the difference in relevance proportion does not impact the effectiveness of this query design method. To conclude, our results indicate that query expansion based on retrieval feedback produces highly significant performance improvements for the MEDLINE database.

CONCLUSIONS

Query expansion strategies are needed to improve users' original queries intended for searching the MEDLINE database. This research, in combination with the results obtained in previous papers [11, 12], indicates that given the current state of art, straight forward statistical and feedback methods when combined with the SMART system's flexible index term weighting options, continue to offer completely viable and effective methods for improving retrieval in MEDLINE.

Further research in this area may be conducted for example, by investigating alternative feedback mechanisms such as Rocchio's [14] method. A second direction of research is to test this method on the larger OHSUMED [18] database. This larger investigation, recently completed, once again indicates the effectiveness of the method. More specifically, when MeSH terms are added to the original OHSUMED queries via retrieval feedback, the 11–AvgP score improves significantly by 9.7% ($p<0.01$) [19]. It is hoped that the combined evidence will provide the encouragement to utilize and further test this feedback approach with other health care databases.

Acknowledgments

The author thanks Professor Hersh for sharing the MEDLINE dataset.

References

1. Hersh W, Hickam D, Haynes R, McKibbon K. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association* 1994;1(1):51–60.

[†] Note that a variety of parameter settings were tried as described in [12].

2. Aronson A, Rindflesch T, Browne A. Exploiting a large thesaurus for information retrieval. *Proceedings of the RIAO Conference (RIAO 94)* 1994;197:216.

3. Rindflesch T, Aronson A. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. *Proceedings of the 18th Symposium on Computer Applications for Medical Care (SCAMC 94).* 1994;240–244.

4. Yang Y. Expert Network: Effective and efficient learning from human decisions in text categorization and retrieval. *Proceedings of the 17th International Conference on Research and Development in Information Retrieval (SIGIR 94* 1994;13–22.

5. Yang Y, Chute C. Words or concepts: the features of indexing units and their optimal use in information retrieval. *Proceedings of the 17th Symposium on Computer Applications for Medical Care (SCAMC 93)* 1993;685–689.

6. Yang Y, Chute C. An application of least squares fit mapping to text. *Proceedings of the 26th International Conference on Research and Development in Information Retrieval (SIGIR 93)* 1993;281–290.

7. Crouch C, Yang B. Experiments in automatic statistical thesaurus construction. *Proceedings of the 15th International Conference on Research and Development in Information Retrieval (SIGIR 92)* 1992;77–88.

8. Jing Y, Croft W. An association thesaurus for information retrieval. *Proceedings of the RIAO Conference (RIAO 94)* 1994;146–160.

9. National Library of Medicine. Unified Medical Language System (UMLS) Knowledge Sources, 5th experimental edition. MD:NLM, 1994.

10. Elkin P, Cimino J, Lowe H, Aronow D, Payne T, Pincetl P, Barnett G. Mapping to MeSH. *Proceedings of the 12th Symposium on Computer Applications for Medical Care (SCAMC 88)* 1988;185–190.

11. Srinivasan P. Query expansion and MEDLINE. *Information Processing and Management* 1996;32(4),431–443.

12. Srinivasan P. Retrieval feedback in MEDLINE. *Journal of the American Medical Informatics Association* 1996;3(2),157–167.

13. Ide E. New experiments in relevance feedback. In: Salton G, ed. The SMART Retrieval System–Experiments in Automatic Document Processing. NJ:Prentice Hall,1971:337–54.

14. Rocchio J. Relevance feedback in information retrieval. In: Salton G, ed. The SMART Retrieval System–Experiments in Automatic Document Processing. NJ:Prentice Hall,1971:68–73.

15. Harman D. Overview of the third Text REtrieval Conference (TREC–3). *Proceedings of the Third Text REtrieval Conference (TREC 3)* 1994;1–19.

16. Buckley C. Implementation of the SMART information retrieval system. Technical Report TR85-686, Department of Computer Science, Cornell University, Ithaca, NY 14853, May 1985.

17. Salton G, ed. The SMART Retrieval System–Experiments in Automatic Document Processing. NJ: Prentice Hall, 1971.

18. Hersh W, Buckley C, Leone T, Hickam D. OHSUMED: An interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th International Conference on Research and Development in Information Retrieval (SIGIR 94)* 1994;192–200.

19. Srinivasan, P. Optimal Document Indexing Vocabulary for MEDLINE. To appear in *Information Processing and Management.* 1996